A study on various aspects of Hate Speech Detection on Social Media using Deep Learning

Synopsis for the Doctoral Degree in Computer Science and Engineering

Submitted By

Rajkumar Vamanrao Panchal

Under the Guidance of

Dr. Megha Venkatesh Jonnalagedda
Associate Professor
Department of Information Technology
Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

Submitted to
Research and Recognition Committee

SWAMI RAMANAND TEERTH MARATHWADA UNIVERSITY
NANDED – 431606, Maharashtra State, India

Contents

1. Introduction ---------- 3

2. Literature Review ---------- 5

3. Objectives ---------- 8

3.1 Expected Outcomes ---------- 8

4. Research methodology ---------- 9

5. Importance of study (applicability to society) ---------- 11

6. Proposed work plan ---------- 12

7. References ---------- 13

1. Introduction

In today's age, social media has become a platform for people to share their thoughts and emotions. Be it a happy or a sad moment people these days tend to turn to social media to express themselves. Many celebrities, influencers and organizations share their thoughts and happenings via social media. This creates huge amount of data and spreading of information at a rapid pace. Some people provide positive and some provide negative comments to these post. The downside of these sites is they do not restrict users to express their thoughts (comments) on posts, due to which many people comment in disparage language. This abusive use of language may hurt someone's feelings or it may create instability in society. So, it is the responsibility of these sites to detect objectionable content before it spreads and results into adverse effects.

It is tedious work to identify textual hate speech manually, due to the high volume of data. So, there must be some way to detect this hate speech automatically. The automatic hate speech detection is also a challenging task due to variation in language, multiple meaning of single statement, difference in syntactic and semantic variation, etc. Specifically in India multiple languages are used. In such multilingual population it becomes challenging to identify hate speech, because many times people use mixed language to share their thoughts or comments.

Now a days great amount of research is going on for automatic identification of hate speech. Various researchers have tried to develop systems for automatic identification of hate speech. There is a lot of research focused on identifying hate speech in English language. Small amount of research is done on hate speech detection in other languages. In India there are 22 languages declared as official languages excluding English. In all there are 780 languages spoken in India, India has the second highest number of languages, as per the People's Linguistic Survey of India. In such vast variety of language variation, automatic detection of hate speech in any single language is challenging a task. Very little research is focused on automatic hate speech detection in Indian languages compared to English.

Moreover in India many times mixed language is used in social media. This mixing of language increases the complexity of automatic hate speech detection. People provide comments by mixing languages like Hindi-English, Hindi-Marathi, Tamil-English and so on. Users of social media write their thoughts, comments in Hindi, Marathi and various other languages using Devnagari script and also for English sentences in Devnagari script. Sometime they write Hindi language in English using Latin script, in such situation users of social media may use different spelling for the same word, which increase the complexity and make the hate speech detection task more challenging.

Automatic hate speech detection is performed by using various techniques. Among these techniques, the system developed using Machine learning and Deep learning become more popular now a days as it provides good result and performance.

Proposed research work try to address some key challenges in hate speech detection and attempt to automatically detect it. Various aspects on hate speech would be studied for detecting hate speech automatically.

2. Literature Review

Hate speech is a form of verbal or non-verbal communication expressing prejudice and aggression. It is defined as an act of belittling a person or community based on their gender, age, sexual orientation, race, religion, nationality, ethnicity etc., [1]. To perform classification feature extraction must be performed and for this purpose there are various ways for feature representation. Following are some feature representation for hate speech detection [2].

1) Dictionaries and Lexicons
2) Character n-grams
3) Bag-of-words (BOW)
4) Term Frequency-Inverse Document Frequency (TF-IDF)
5) Latent Dirichlet Allocation (LDA)
6) Word embedding and Word2Vec
7) FastText
8) Glove

For performing classification of hate speech machine learning or deep learning algorithm can be used.

1) Machine Learning Algorithm: Bayesian Logistic Regression (BLR)[3], Support Vector Machine (SVM)[3][4], Random Forest (RF) [3][4], Logistic Regression[6], Decision Tree [6]
2) Deep Learning Algorithm: Convolutional Neural Network (CNN)[7], Long Short-Term Memory(LSTM)[7], Bidirectional LSTM(BiLSTM)[7][11], C-BiGRU[8] and Bidirectional Encoder Representations from Transformers(BERT)[9], DistilBERT[6], MuRIL[6], mBERT [10] .

- Machine Learning Algorithm:-

In the early research on hate speech detection, the use of machine learning algorithms was preferred. Any machine learning or deep learning algorithm accepts only number as input, so textual data must be converted into vectors, so that it can be provide as input to classifier. Using FastText, Word2Vec/Doc2Vec word embedding and feature extraction performed. After preprocessing machine learning algorithm were used for detection of hate speech text in Hindi-English Code-mixed Data. Among SVM-linear, SVM-RBF (Radial Basis Function), Random Forest (RF), performance of SVM-RBF was better with accuracy 0.85 [1].

A data from Twitter used to perform hate speech detection, each tweet transformed to vector, before doing this removal of stop words, nonalphanumeric character accomplished. All the features present in dataset were represented using Bag-of-words (BOW) and N-grams, then classifier model used to identify cyber hate spreading online. Mainly three classifier model used for classification purpose, Bayesian Logistic Regression (BLR), Random Forest Decision Tree

(RFTD), Support Vector Machine (SVM). Overall 0.95 precision, recall and F-Measure reported [3].

In Natural Language Understanding (NLU) feature extraction is an important step, Word2Vec performs better feature representation as compared to FastText for the task of classification in NLU[4].

Aditya et.al [5] collected around 4575 Hindi-English tweets, the data was preprocessed and emoticon count, punctuation count, character n-gram, word n-gram, word2vec features were extracted. Using chi square dimensionality reduction technique the size of extracted feature reduced. Using Random Forest and SVM classification algorithm hate speech detection performed which gave accuracy of 66.7% and 71.7% respectively.

- Deep Learning Algorithm:-

Machine Learning algorithms require feature selection for hate speech detection on contrary deep learning algorithm automatically select feature from input data.

Santosh et al. [11] performed hate speech detection on social media text using two deep learning algorithms. Sub-word level LSTM model and Hierarchical LSTM model with attention based phonemic sub-word. In sub-word level LSTM input data is converted into sub-word level representation which provide feature map of input data. Through max pooling this feature map is provided to LSTM which results into detection of hate speech. Variations in spelling can gives rise to out-of-vocabulary words, to tackle this problem words are segmented into phonemic sub words, which is used as input to Hierarchical LSTM model. Hierarchical LSTM model uses embedding layer to convert sentences into low dimension representation. Which then given to syllable encoder. This encoded data is used as input to next layer that is word encoder. In the next layer that is word attention important words are attended. Finally output layer detect hate speech. The overall accuracy of Sub-word level LSTM model and Hierarchical LSTM model is 69.8% and 66.6% respectively.

Deep learning methods like 1D-CNN, LSTM and BiLSTM is used for hate speech detection in Marathi language [7]. The word embedding method like FastText combined with deep learning methods. The dataset provided by Hasoc-2021 is used for training and testing. 1D CNN achieved an accuracy of 83%. Apart from this transformer based embedding's like Roberta-mr, IndicBERT, m-BERT used for hate speech detection. In transformer based approaches IndicBERT performed the best with an accuracy of 88%.

RNN is also used for hate speech detection. Hate speech and offensive content detection in Hindi language composed of CNN together with bidirectional RNN [8]. The C-BiGRU uses the data provided by HASOC 2021 workshop. In preprocessing step emoticons, URL etc. removed from tweets. This preprocessed tweets are fixed to a sequence length of 150 tokens, which is provided to convolution layer and then to GRU. The accuracy of the C-BiGRU is 77.48% and F1 score of 75.04%.

There are various word embedding techniques which are used to convert word into vector. This embedding store meaningful information about words such as semantic relationship,

syntactic relationship and contextual information. The word embedding techniques mostly used for hate speech detection are Word2Vec, Glove, FastText, BERT. Among these, variants of BERT provided prominent result for detection of hate speech. The accuracy of DistilBERT and MuRIL is 0.76 and 0.79 respectively [6].

For hate speech detection in code mixed Hindi, hate speech is categorized hate speech into 34 categories. Among these 34 categories 28 were for single Hindi Language and remaining 6 were for code mixed hate speech. To detect hate speech multilingual transformer model (m-bert) used [10]. This work also identified hateful speech with missing offensive words and missing names of the protected groups against which the hate is being directed.

Hate speech contents created lot of violence and with the help of social media these hate speech spread rapidly in society. Many time social media unable to detect hate text and face lot of criticism due to increasing offensive content in their sites. Multiple concepts related to hate speech used in NLP such as: abusive language, profanity, flaming, discrimination, toxic language or comment. Multiple approaches in NLP used to detect hate speech. Using CNN tweets in dataset classified into hate, offensive and neither [12] for English language. Overall 91% accuracy achieved through CNN.

Hindi is complex language when it written using roman scripts, due to variations in grammar, spelling and translation makes the hate and offensive content detection a difficult task. In India rising use of regional language in social media increased the need for researcher to find hate speech [13].

The dataset for Hinglish language introduced with 3000data, lebeled as Non-offensive, Abusive and Hate-inducing [14]. The deep learning algorithm CNN and LSTM is used for classification of offensive tweets. To achieve better result different word embedding technique such as word2vec, FastText and Glove are used as feature. Using transfer learning approach along with deep learning model 0.89 F1 score achieved for Hinglish language.

3. Objectives

Proposed research work objectives are enlisted below.

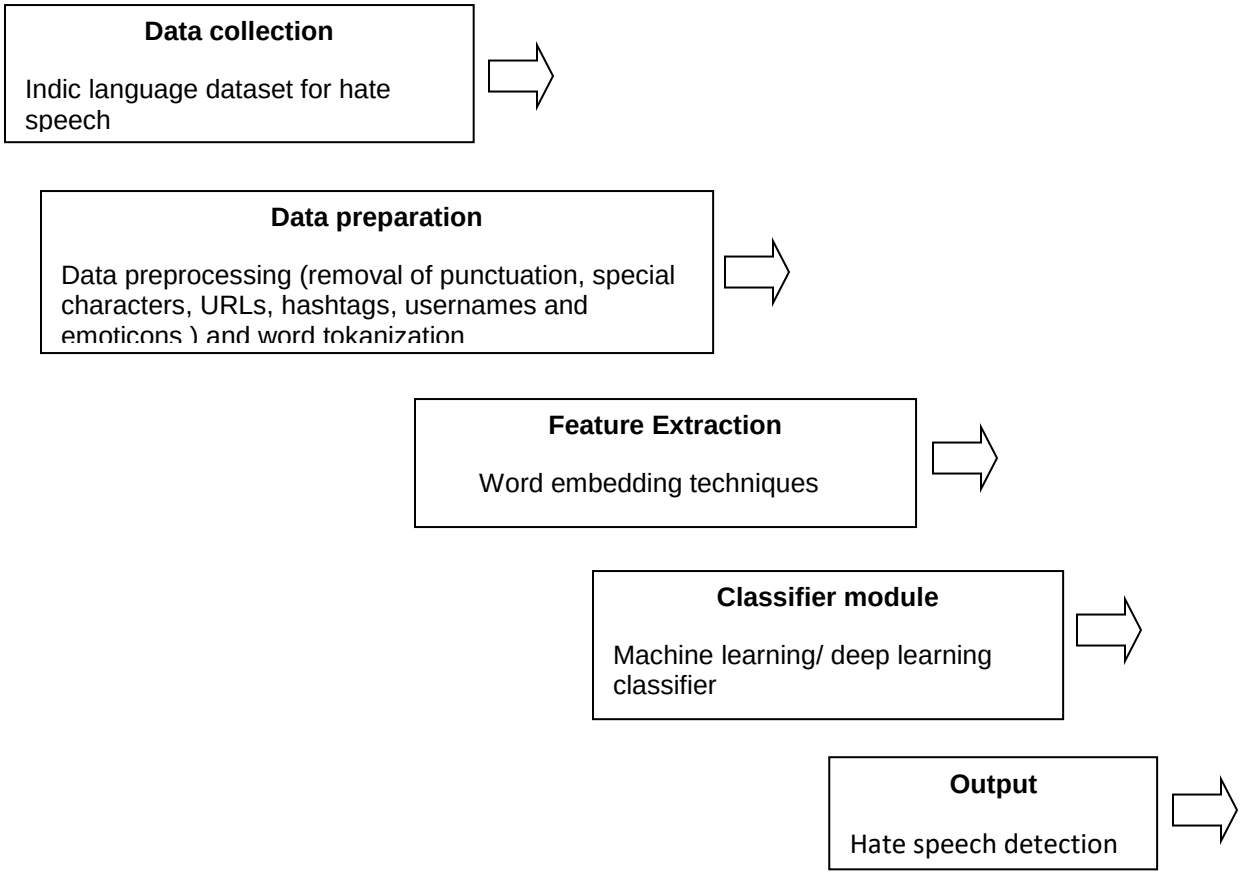
- Study currently available hate speech detection technique for social media text
- Explore challenges in NLP for hate speech detection
- Analyze available automatic hate speech detection techniques for Indic language
- Develop automatic hate speech detection system for Indian language
- Classify text on social media into hate, offensive, non-hate categories

3.1 Expected Outcomes

- Proposed research work will provide a systematic approach to analyze available Hate speech detection techniques/ methodologies.
- It will aim to address and overcome the challenges and disadvantages in present research work.
- Proposed research work is focused on classifying hate speech detection of Indic languages.

4. Research methodology

Proposed research methodology is as mentioned below. Fig. 1 shows the flow chart of the proposed methodology. Data collection of the Indic language for hate speech. Data preprocessing and hate speech detection/classification.


Data collection
Indic language dataset for hate speech
Data preparation
Data preprocessing (removal of punctuation, special characters, URLs, hashtags, usernames and emoticons) and word tokanization
Feature Extraction
Word embedding techniques
Classifier module
Machine learning/ deep learning classifier
Output
Hate speech detection

In proposed methodology available hate speech detection dataset for Indian language will be used.

- **Input Dataset:** To work on Indian language for hate speech detection, available standard dataset will be used. Dataset provided by HASOC 2023 or CONSTRAINT – 2021 or MahaBERT on HuggingFace or English will be used for research work.
- **Data preparation module:** As dataset contains text from social media like twitter, Facebook etc. Dataset may contains punctuation, special characters, URLs, hashtags, usernames, emoticons, and extra white spaces, removal of these things is necessary. In data preprocessing, removal of punctuation, special characters, URLs, hashtags, usernames, emoticons, and extra white spaces performed. Then in word tokenization tokens for word are created.

- **Feature Extraction:** To detect hate speech effectively feature extraction is necessary. In this module features from inputs are extracted. For feature extraction there are various techniques available such as
  1. Dictionaries and Lexicons
  2. Character n-grams
  3. Bag-of-words (BOW)
  4. Term Frequency-Inverse Document Frequency (TF-IDF)
  5. Latent Dirichlet Allocation (LDA)
  6. Word embedding and Word2Vec
  7. FastText
  8. Glove

  Once features are extracted words are converted into vector. Word embedding store meaningful information about words such as semantic relationship, syntactic relationship and contextual information.

- **Classifier module:** The output of feature extraction module is provided as input to classifier module. In classifier module, classification of text will be performed using machine learning or deep learning algorithm.
- **Output module:** The output of classifier is hate speech detection.

5. Importance of study (applicability to society)

The use of social media created enormous text, this large amount of text is about someone's thoughts, daily happening, information about event, comments and lot more. The increased use of social media helped in globalization but the downside of this is some people uses this platform to create instability in society, hurting someone's feelings, belittling a person or community based on their gender, age, sexual orientation, race, religion, nationality, ethnicity etc. To avoid this there is need of automatic identification of hate speech. This detection of hate speech prevent society from being unstable, avoid person from feeling psychologically negative, avoid person from being criminal and avoid person suicide.

Proposed research work will try to address challenges in detecting cyber hate. This research work also come up with various aspects where challenges from existing models/ technologies would be addressed in efficient way using machine learning/deep learning technique.

6. Proposed work plan

#	Tasks	Timelines	Dependencies - if any	Completion criteria	Status
1	Literature review	March 2023 to August 2023	None	Synopsis preparation	Completed
2	Synopsis review and revisions with guide	August and September 2023	Dependent on task 1	Approved synopsis for RRC	planned
3	RRC research proposal presentation and approval	September to December 2023	Dependent on task 2	Research title approval letter from RRC	planned
4	Detailed study w.r.t. research proposal objectives	January 2024 to April 2024	None	Survey paper	Planned
5	Identification of gaps in current approaches related to 1st key HSD challenge	May 2024 to August 2024	Dependent on task 4	Probable solution identification w.r.t. gaps identified	Planned
6	Validation of proposed approach for 1st key HSD challenge	September 2024 to December 2024	Dependent on task 5	Publication draft creation	Planned
7	Review with guide, feedback incorporation	January 2025 to February 2025	Dependent on task 6	Publication in peer reviewed journal (Research center approved)	Planned
8	Identification of gaps in current approaches related to 2nd key HSD challenge	February 2025 to May 2025	Dependent on task 7	Probable solution identification w.r.t. gaps identified	Planned
9	Validation of proposed approach for 2nd key HSD challenge	June 2025 to August 2025	Dependent on task 8	Publication draft creation	Planned
10	Review with guide, feedback incorporation	August 2025 to September 2025	Dependent on task 9	Publication in peer reviewed journal (Research center approved)	Planned
11	Identification of gaps in current approaches related to 3rd key HSD challenge	October 2025 to December 2025	Dependent on task 10	Probable solution identification w.r.t. gaps identified	Planned
12	Validation of proposed approach for 3rd key HSD challenge	January 2026 to May 2026	Dependent on task 11	Publication draft creation	Planned
13	Review with guide, feedback incorporation	May 2026 to June 2026	Dependent on task 12	Publication in peer reviewed journal (Research center approved)	Planned
14	Thesis first draft creation	July 2026 to October 2026	Dependent on all above tasks	1st Draft of thesis ready for review	Planned
15	Thesis draft review with guide, revisions	November 2026 to December 2026	Dependent on all above tasks	Thesis ready (post guide's approval) for next steps	Planned

7. References

[1] Sreelakshmi k,, Premjith B,and Soman K.P. (2020) “Detection of Hate Speech Text in Hindi-English Code-mixed Data.” Procedia Computer Science 171 (2020) 737–744

[2] Al-Hassan, Areej, and Hmood Al-Dossari. (2019) “Detection of hate speech in social networks :a survey on multilingual corpus.” Computer Science Information Technology (CSIT) 9 (2) :8.

[3] P. Burnap and M. L. Williams, “Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making,” Policy & Internet, vol. 7,
no. 2, pp. 223–242, 2015.

[4] Premjith, Bhavukam, Kutti Padannayil Soman, and Prabaharan Poornachandran. (2019) “Amrita CEN@ FACT: Factuality Identification in Spanish Text.” Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain

[5] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. (2018) “ A dataset of hindi-english code-mixed social media text for hate speech detection.” Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media :36-41

[6] Bansod, Pranjali Prakash, "Hate Speech Detection in Hindi" (2023). Master's Projects. 1265. DOI: https://doi.org/10.31979/etd.yc74-7qas ,https://scholarworks.sjsu.edu/etd_projects/1265

[7] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, ‘‘L3cube-mahahate: A tweet-based marathi hate speech detection dataset and bert models,’’ arXiv preprint arXiv:2203.13778, 2022.

[8] Sudharsana Kannan and Jelena Mitrović(2021),” Hatespeech and Offensive Content Detection in Hindi Language using C-BiGRU.” Forum for Information Retrieval Evaluation, FIRE 2021, Gandhinagar, India, December 13-17, 2021, p. 209-216.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ‘‘Bert: Pre-training of deep bidirectional transformers for language understanding,’’ arXiv preprint arXiv:1810.04805, 2018.

[10] M. Das, P. Saha, B. Mathew, and A. Mukherjee, ‘‘Hatecheckhin: Evaluating hindi hate speech detection models,’’ arXiv preprint arXiv:2205.00328, 2022.

[11] Santosh, T. Y. S. S., and K. V. S. Aravind. (2019) “Hate Speech Detection in Hindi-English Code-Mixed Social Media Text.” Proceedings of the ACM India Joint International Conference on Data Science and Management of Data ACM : 310-313.

[12] Biere, Shanita, Sandjai Bhulai, and Master Business Analytics. (2018) “Hate Speech Detection Using Natural Language Processing Techniques”. Diss. Master’s dissertation, Vrije Universiteit Amsterdam.

[13] Bali, Kalika, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. (2014) " "I am borrowing ya mixing?" An Analysis of English-Hindi Code Mixing in Facebook." Proceedings of the First Workshop on Computational Approaches to Code Switching : 116-126

[14] Mathur, Puneet, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. (2018) "Did you offend me? classification of offensive tweets in hinglish language." Proceedings of the 2nd Workshop on Abusive Language Online (ALW2) :138-148

Research Scholar

Rajkumar Vamanrao Panchal

Research Guide

Dr. Megha Venkatesh Jonnalagedda

Associate Professor, Department of Information Technology

Shri Guru Gobind Singhji Institute of Engineering and Technology, Nanded

Date: 28th August 2023.

Place: Nanded, Maharashtra, India